INTERPRETING DATA FOR PART II PATHOLOGY

Caroline Trotter
clt56@cam.ac.uk

PART II PATHOLOGY

Pathology is a **quantitative** discipline

A range of different methods are used to describe and measure biological phenomenon

Different statistical methods for analysing data may be >< important in the different courses but there are some important underlying principles

All students need to be able to interpret scientific literature

Single Subject students will need to analyse the data generated in their project

LEARNING OBJECTIVES

During this lecture you will learn:

- to recognise and define measures of central tendency, variability and range
- the principles of sampling methods and hypothesis testing
- to interpret p-values and confidence intervals
- to appreciate the difference between statistical and biological significance

VARIABILITY

Most things you'll come across in Part II Pathology (and indeed most biological phenomena you can think of) will display variability

Frequency distributions express this variability and are summarised by measures of central tendency and spread.

mean

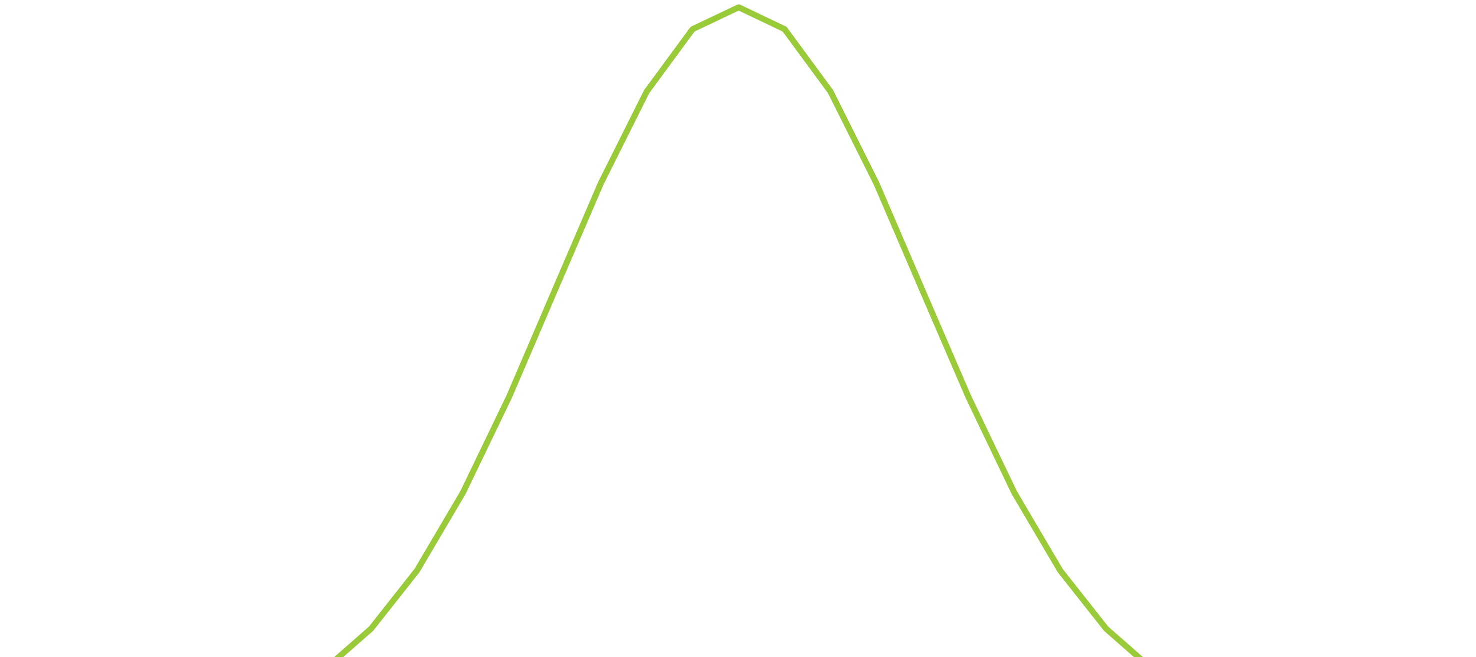
median

mode

range

interquartile range

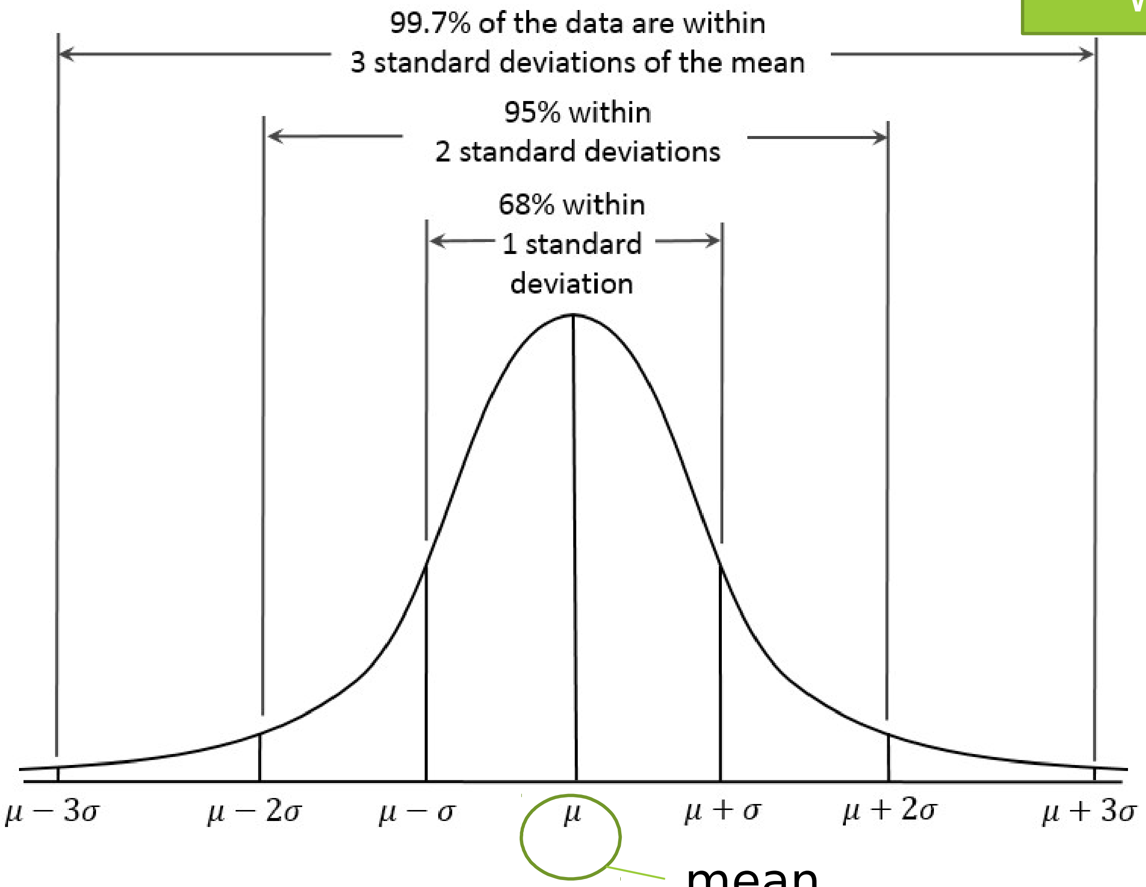
standard deviation



NORMAL DISTRIBUT

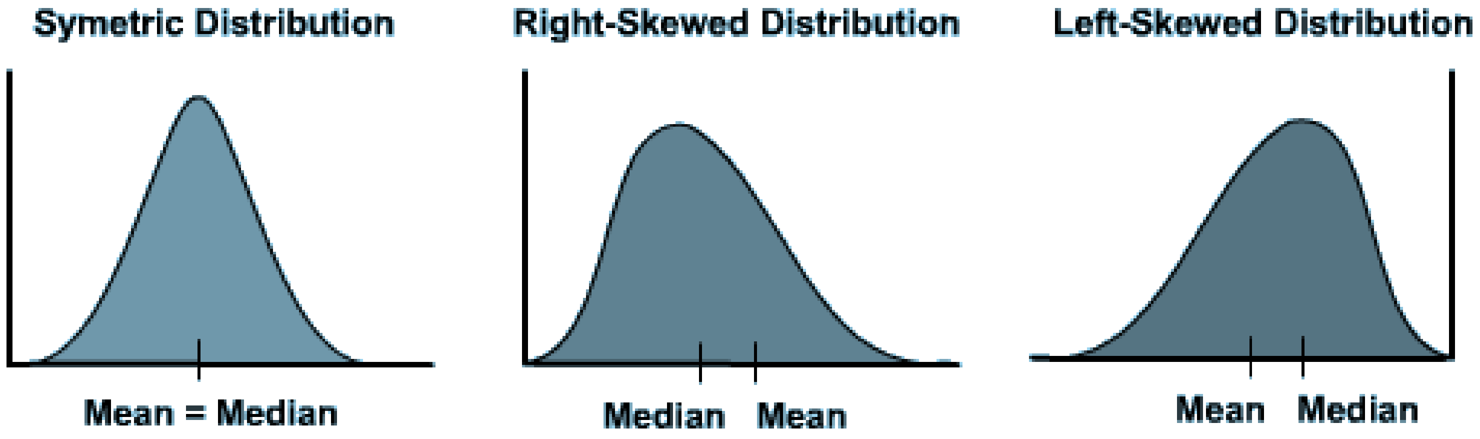
Small standard deviation tall, narrow curve

Large standard deviation short, wide curve



mean

SKEWED DATA


Symetric Distribution
Mean = Median
Right-Skewed Distribution
Median
Mean
Left-Skewed Distribution
Mean
Median

CONTINUOUS VS DISCRETE VARIABLES

Continuous variable (numerical)

- can take any value

- e.g. height, weight, antibody concentration

Discrete variable (categorical)

- can only be one of a set number of categories

- binary outcome: e.g. disease or not disease

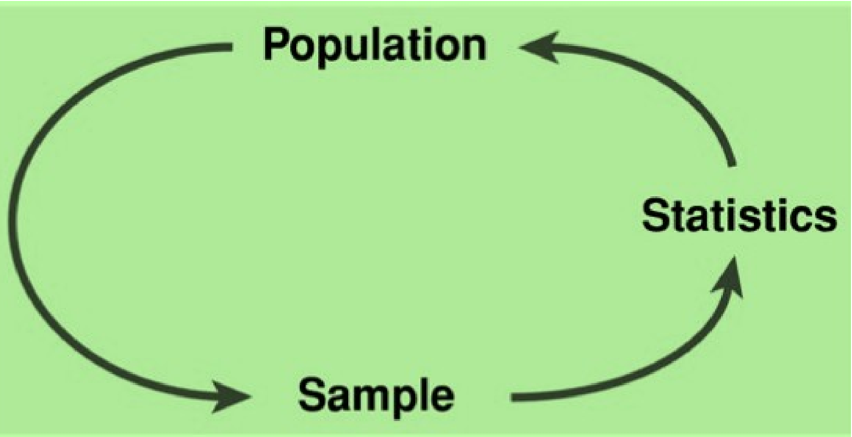
- unordered categorical: e.g. marital status

- ordered categorical: e.g. degree class

INFERENCES FROM A SAMPLE

Most of the time it is not possible to measure the whole population

If a **representative** sample is obtained, inferences can be made about the population of interest



SAMPLING VARIATION

Estimates obtained from a sample of the population suffer from sampling variation

It is important to take this into account - this is often don’t by reporting confidence intervals

SAMPLING DISTRIBUTION

If one repeatedly took samples from the target population, you would get a different mean each time. However:

- provided the sample size is large enough the sample means have an approximately Normal distribution (even if the population distribution is not normal)

- the mean of the sample distribution is equal to the population mean

- the standard deviation of the sampling distribution (the standard error) depends on the variation in the population and the size of the samples

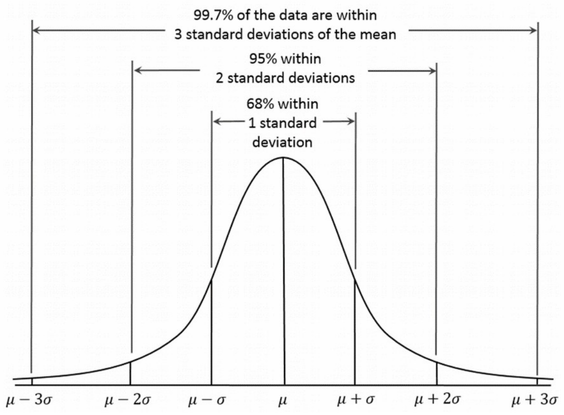
CONFIDENCE INTERVALS

Because we know that the distribution of sample means is Normal, we can say that 95% of the individual sample means are within 1.96 standard errors of the mean of this distribution

Transposing this…

95% of the time the true population mean is within 1.96SEs of the observed sample mean

If we draw several, independent random samples from the same population and calculate confidence intervals then on average 19/20 (95%) of the CI would contain the true population mean



COMPARISONS BETWEEN GROUPS

Often we are interested in comparing quantities two or more groups rather than estimating single quantities.

Confidence intervals may still be useful as they represent the effect size and the degree of accuracy that we have in the estimated value

E.g. 1.80 m (95% CI 1.75, 1.85m)

1.65m (95% CI 1.61, 1.69m)

but it may be more useful to formulate a **hypothesis test**

HYPOTHESIS TESTING

Null hypothesis (H_0)- states no difference / association

E.g. there is no difference in the mean height of Cambridge and Oxford students

As with confidence intervals, we can use a sample (and its distribution) to tell us about a population

We calculate a test statistic, based on the data, and then compare this to its sampling distribution.

This will give a p-value – i.e. the chance of seeing a sample estimate at least as different from the null hypothesis as the one we observe

DIFFERENCE BETWEEN MEANS

We want to compare the means between two groups of normally distributed individuals. To

do this we specify a null hypothesis and an alternative hypothesis

and try to build evidence against the null.

H0 : 1 = 2 vs. H1 : 1 6= 2

This is an example of a two-sided hypothesis. The null is that the

population means are equal in both groups, and the alternative is

that they are not equal.

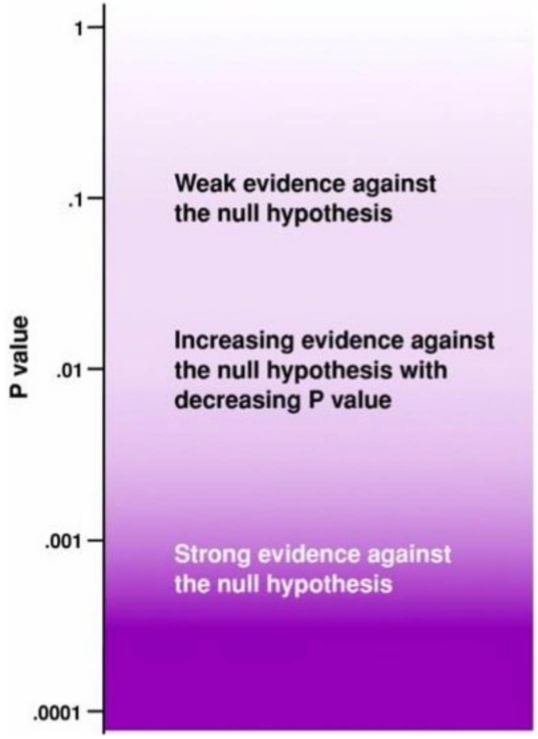
To test this we calculate a test statistic, based on the data, and

then compare this to its sampling distribution.

INTERPRETING P-VALUES

p<0.05 is arbitrary and although widely used this should be discouraged!

(See article by Sterne & Davey Smith in suggested reading)


1
.1
.01
.001
.0001
P value
Weak evidence against the null hypothesis
Increasing evidence against the null hypothesis with decreasing P value
Strong evidence against the null hypothesis

WHAT KIND OF TEST?

The type of hypothesis test depends on the question and the data

Some examples given in the handout.

- Comparing 2 means: t-test / Z-test
- Comparing means between multiple groups: ANOVA
- Comparing locations between multiple (or two) groups:

Kruskal-Wallis (Mann-Whitney).

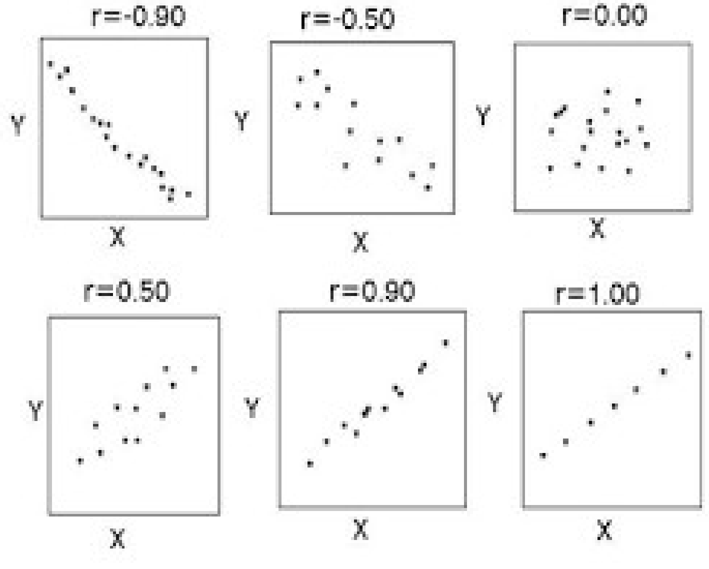
- Looking for a linear association between two continuous/discrete variables: correlation coefficient.
- Looking for an association between two categorical variables: chi-squared test.

LINEAR REGRESSION AND CORRELATION

For two continuous (numerical) variables:

Linear regression can be used to estimate the best-fitting straight line to describe the association between the variables

Correlation coefficient can be estimated to examine the strength of linear associat



SIMPLE LINEAR REGRESSION

If the model fit is **reasonable,** we can make **predictions**

(Be careful no to extrapolate beyond the range of the data since the linear assumption may not hold)

The framework can be extended to non-Normal data(e.g. generalised linear models), non-linear regression terms and multiple response variables (multivariate regression) and multiple explanatory variables (multiple regression)

F statistic and p-value examine whether the independent variables reliably predict the dependent variable (hypothesis test)

R^2 provides an estimate of the strength of the relationship between your model and the response variable

BIOLOGICAL SIGNIFICANCE

In every case, consider not just the statistical significance of the results, but also the biological significance

Anti-tetanus antibodies in children with and without prophylactic paracetamol at vaccine administration

N	Protective cut-off	GMCs
206	100%; (98·2–100)	1·639 (1·474–1·822)
225	100%; (98·4–100)	2·669 (2·434–2·927)*

VISUALISING DATA

[Figures removed for copyright purposes]

There are a huge variety of ways of presenting data visually

This can really help in understanding and interpreting data

Think about how to do this for your own project

What works well in papers you have read?

INTERPRETING DATA IN THE LITERATURE

Look at the NUMBERS of subjects, experiments …

Are the data presented visually?

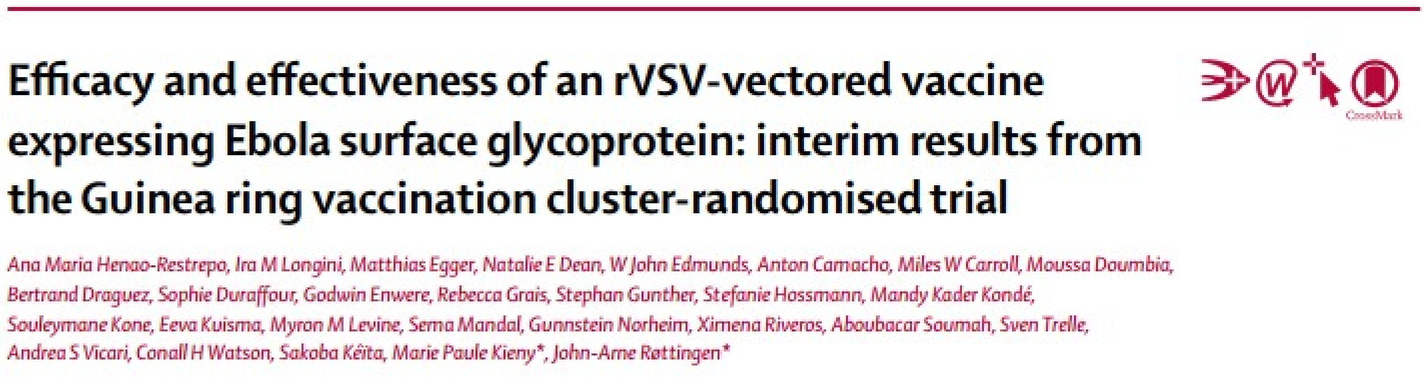
Are the statistical tests described? (are they appropriate?)

Interpret the 95% CI and p-values

What is the biological significance?

Do you agree with the authors’ conclusions?

EXAMPLE



Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial

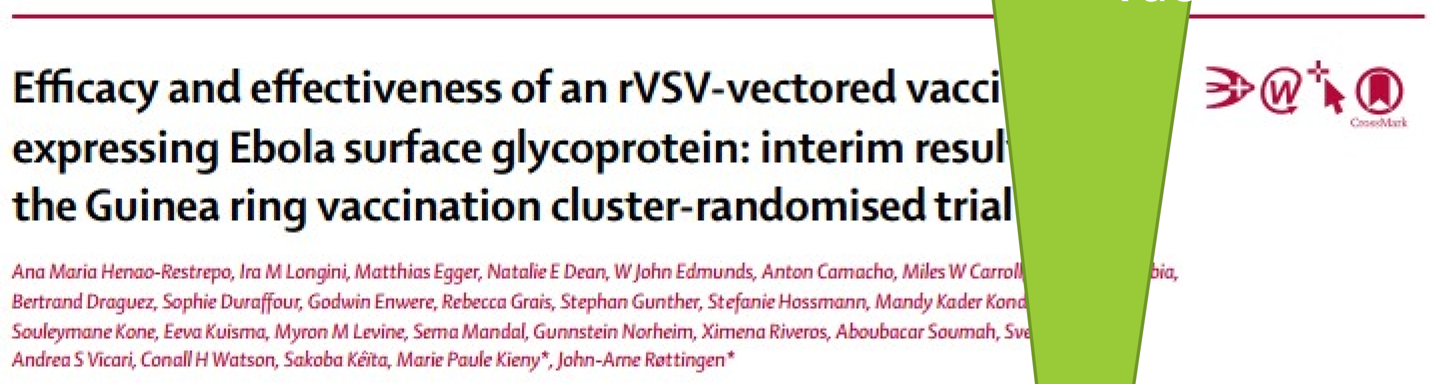
Ana Maria Henao-Restrepo, Ira M Longini, Matthias Egger, Natalie E Dean, W John Edmunds, Anton Camacho, Miles W Carroll, Moussa Doumbia, Bertrand Draguez, Sophie Duraffour, Godwin Enwere, Rebecca Grais, Stephan Gunther, Stefanie Hossmann, Mandy Kader Kondé, Souleymane Kone, Eeva Kuisma, Myron M Levine, Sema Mandal, Gunnstein Norheim, Ximena Riveros, Aboubacar Soumah, Sven Trelle, Andrea S Vicari, Conall H Watson, Sakoba Kéïta, Marie Paule Kieny, John-Arne Røttingen**

90 clusters with a total population of 7651 people were included in the analysis.

In the immediate vaccination group, there were no cases of Ebola virus disease with symptom onset at least 10 days after randomisation, whereas in the delayed vaccination group there were 16 cases of Ebola virus disease from seven clusters, showing a vaccine efficacy of 100% (95% CI 74·7-100·0; p=0·0036).

EXAMPLE

Vaccine efficacy (VE) =
(ARU - ARV)/ARU
(x 100).
[ARU= attack rate in
unvaccinated
ARV= attack rate in
vaccinated]



Efficacy and effectiveness of an rVSV-vectored vacci
expressing Ebola surface glycoprotein: interim resul
the Guinea ring vaccination cluster-randomised trial

Ana Maria Henao-Restrepo, Ira M Longini, Matthias Egger, Natalie E Dean, W John Edmunds, Anton Camacho, Miles W Carroll bia,
Bertrand Draguez, Sophie Duraffour, Godwin Enwere, Rebecca Grais, Stephan Gunther, Stefanie Hossmann, Mandy Kader Kond
Souleymane Kone, Eeva Kuisma, Myron M Levine, Sema Mandal, Gunnstein Norheim, Ximena Riveros, Aboubacar Soumah, Sve
Andrea S Vicari, Conall H Watson, Sakoba Kéïta, Marie Paule Kieny*, John-Arne Røttingen*

90 clusters with a total population of 7651 people were included in the analysis.

In the immediate vaccination group, there were **no cases** of Ebola virus disease with symptom onset at least 10 days after randomisation, whereas in the delayed vaccination group there were **16 cases** of Ebola virus disease from seven clusters, showing a vaccine efficacy of **100% (95% CI 74·7-100·0; p=0·0036).**

FURTHER READING

McKinley TJ. Part II Pathology: Basic Statistics. (Handout)

Kirkwood BR, Sterne JAC (2003) Essential Medical Statistics. 2nd edition. Blackwell Science Ltd.

Sterne JAC, Davey Smith G (2001) Sifting the evidence: what’s wrong with significance tests? BMJ 322: 266-31

Ben-Shlomo Y, Brookes ST, Hickman M (2013). Epidemiology, evidence-based medicine and public health. Wiley-Blackwell.